1. **Provide your team background and organization description (if applicable).**
   - None, I am a student doing ML.

2. **Explain why you participated in the Allergen Chip challenge.**
   - Answer - My participation in the Allergen Chip Challenge was motivated by a desire to delve into the complexities of integrating different types of data for machine learning applications. The challenge presented a unique opportunity to work with both tabular data and image data, although my final solution focused primarily on the former.
   - The experience allowed me to deepen my understanding of tabular data and the importance of feature engineering in machine learning. I was able to explore various techniques for processing and analysing this type of data, and I gained valuable insights into how it can be used to solve complex problems in the field of healthcare.

3. **Describe how you built your winning model and elaborate on the technical and modeling choices you made.**
   - Preprocessing - I utilized Multilabel Stratified K-Fold cross-validation due to the multi-target nature of the problem. This ensured a balanced representation of the targets in each fold. Additionally, I engaged in feature engineering, creating new features to capture complex patterns in the data. Specifically, I calculated row-wise statistics such as mean, standard deviation, product, sum, and median across the metadata columns. I also created a feature that represented the multiplication of the row mean and standard deviation. These engineered features provided a more comprehensive representation of the data, capturing potential interactions and dependencies between the original features. By incorporating these features into the model, I was able to significantly improve the model's F1 score, demonstrating their effectiveness in enhancing the model's predictive capabilities.

   - In the model building phase, I utilized an XGBoost model, wrapped in a MultiOutputClassifier, to handle the multi-label nature of the problem. After experimenting with various architectures such as Linear Layers, Catboost, LightGBM, and ensemble methods, a well-tuned XGBoost model was the most effective. The model was configured for binary classification with 950 boosting rounds, a learning rate of 0.06, and a 'colsample_bytree' of 0.5 to prevent overfitting. I employed a 15-fold cross-validation strategy, training the model on all other folds and validating it on the current fold for each iteration.

- In the post-processing phase, I focused on optimizing the decision threshold for each target label to maximize the F1 score. This step was crucial as it significantly improved the model's F1 performance. Instead of using a standard threshold of 0.5 for all targets, I calculated the optimal threshold for each target individually. This was achieved by computing the F1 score for a range of thresholds and selecting the one that yielded the highest score for each target.

- Interestingly, I observed a strong correlation between the number of positive samples in a target and its optimal threshold. As the number of positive samples increased, a higher threshold was required to achieve the best performance. This relationship was visualized in a scatter plot, which clearly showed the trend and the strong correlation of 0.945.

- This approach of finding a unique threshold for each target was a key factor in the performance of my model. It allowed the model to adapt to the unique characteristics of each target, thereby improving its overall performance.

4. **Were the CPU/RAM resources provided in the challenge notebook sufficient from your point of view?**
    - I feel like the ram resources were sufficient, but the CPU resources were not sufficient for fast iteration. One of the major reasons I did not use images and was not able experiment with them enough was because of a lack of a GPU and CPU resources.